DISCUSSION PAPERS
Department of Economics
University of Copenhagen

06-11

Testing Preference Axioms in Discrete Choice experiments: A Reappraisal

Jens Leth Hougaard, Tue Tjur
& Lars Peter Østerdal

Studiestræde 6, DK-1455 Copenhagen K., Denmark
Tel. +45 35 32 30 82 - Fax +45 35 32 30 00
http://www.econ.ku.dk

Testing preference axioms in discrete choice experiments: a reappraisal

Jens Leth Hougaard
Department of Economics
University of Copenhagen

Tue Tjur
The Statistics Group
Copenhagen Business School

Lars Peter Østerdal
Department of Economics
University of Copenhagen

June 2006

Abstract

Recent studies have tested the preference axioms of completeness and transitivity, and have detected other preference phenomena such as unstability, learning- and tiredness effects, ordering effects and dominance, in stated preference discrete choice experiments. However, it has not been explicitly addressed in these studies which preference models are actually being tested, and the connection between the statistical tests performed and the relevant underlying models of respondent behavior has not been explored further. This paper tries to fill that gap. We specifically analyze the meaning and role of the preference axioms and other preference phenomena in the context of stated preference discrete choice experiments, and examine whether or how these can be subject to meaningful (statistical) tests.

JEL classification: B41, C52, D01.

Keywords: Stated preference discrete choice experiments, Completeness, Transitivity, Random utility, Statistical tests.

Correspondence: Lars Peter Østerdal, Department of Economics, University of Copenhagen, Studiestraede 6, DK-1455 Copenhagen K, Denmark. Phone: +45 35 32 35 61. Fax: +45 35 32 30 85. E-mail: lars.p.osterdal@econ.ku.dk

Acknowledgements: We would like to thank Martin Browning, John Cairns, Søren Johansen, and Helgi Tómasson for helpful conversations and comments to a preliminary draft [14]. The usual disclaimers apply.

1 Background and introduction

When researchers and practitioners build statistical models of individual behavior or welfare, they often assume that choices are guided by utility maximization or preferences in accordance with certain basic consistency requirements (sometimes called "axioms").

In the particular context of consumer demand under (non-)linear budget constraints, the utility maximizing hypothesis is, in theory, testable through the *revealed preference axioms*. There is an extensive literature on this topic and classical references include Houthakker [15], Richter [34], Afriat [1], Diewert [7], Varian [47], Matzkin [26], and Matzkin and Richter [24]. More recently, stochastic tests have been developed by Cox [6], Fleissig and Whitney [12] and others.

Another line of the literature has indicated an increased interest in testing the preference axioms of *completeness* and *transitivity*, and detecting other preference phenomena such as *unstability*, *learning- and tiredness effects*, *ordering effects*, *dominance*, etc., in stated preference discrete choice experiments. Recent studies along these lines include Ryan et al. [37], Shiell et al. [41], Carlsson and Martinsson [5], Johnson and Mathews [19], Sælensminde [44][45], Ryan and Bate [36], McIntosh and Ryan [32], San Miguel et al. [38], Scott [39], and Ryan and San Miguel [35].

Stated preference discrete choice experiments are widely used, for example in environmental evaluation, marketing- and transportation studies and for evaluation of health states or health care services. In light of the popularity of such experiments testing its preference foundation is of broad interest. Further it can be motivated by stated preference experiments that often have no real consequences for the respondents involved (and hence very little can be assumed a priori about their behavior). Somewhat surprisingly, though, the second line of literature has not explicitly addressed the question of which preference models that are actually being tested and whether the preference axioms (and other associated preference phenomena) are, in fact,

subject to statistical testing under the given circumstances. Some of these studies refer to the random utility model, with the explicit aim of testing the underlying preference axioms, but the link between the statistical tests performed and the random utility model of respondent behavior is typically not fully explored, and it is questionable whether the proposed tests relate to a validation of this model at all.

This paper is an attempt of a methodological and interpretive reappraisal. We discuss how the preference axioms and other preference phenomena can be specified meaningfully in relation to various discrete choice models, and examine whether and how these can be subject to meaningful (statistical) tests.

Section 2 begins by recalling the fundamentals concerning the role of the axiom of completeness for (tests of) the utility maximizing hypothesis, arguing that, within an abstract decision model, it is a technical assumption that one may, or may not, impose initially, but it has no real behavioral substance. Theoretically, it does play a minor role for the verifiable implications of utility representations, but in the present context it is unlikely to have much relevance, if any. Section 2.1 focusses on recent experimental studies. Tests of the completeness axiom have recently drawn attention although we shall argue that, within the relevant statistical models, it is meaningless to test for completeness since it cannot be disentangled from random error. The basic methodological issues we attempt to raise apply to detection of other associated preference phenomena as well.

In case of repeated choices, discussed in Section 3, preferences are naturally interpreted as choice frequencies (e.g., May [25]) and it becomes impossible to distinguish between "coin-flip" answers (interpreted as incompleteness) and similarity of alternatives (interpreted as indifference). Consequently, tests for completeness cannot be performed within this model. Tests for transitivity can be meaningful in relation to at least one statistical model, and we suggest one way to perform such a test. Section 3.1 discusses

recent experimental tests of the transitivity axioms which seem to be of a very different kind.

In Section 4, we further connect this discussion to the random utility model (McFadden [28]). Within the framework of this more structured model, underlying preference relations are in a sense not only complete but also transitive by construction. But, in Section 4.1, we suggest how various other preference phenomena could be interpreted and detected within extended versions of this model. Section 5 concludes.

2 The completeness axiom and tests of the utility maximizing hypothesis

In experimental studies, tests of the axiom of completeness are sometimes associated with tests of the utility maximizing hypothesis. Hence it seems useful to recall the role of the completeness axiom in abstract utility theory.

Suppose that there is a finite set of alternatives X. A preference relation is a binary relation $\succsim$ on X where we interpret $x \succsim y$, $x, y \in X$, as "x is at least a good as y". From $\succsim$ we define strict preference $\succ$ and indifference $\sim$ in the usual way, i.e. $x \succ y$ if $x \succsim y$ and not $y \succsim x$, and $x \sim y$ if $x \succsim y$ and $y \succsim x$.[1] It is well-known that if $\succsim$ is complete (i.e. $x \succsim y$ or $y \succsim x$ for all $x, y \in X$) and transitive (i.e. $x \succsim y$ and $y \succsim z$ implies $x \succsim z$), then there exists a real-valued function u on X that represents $\succsim$ in the sense that

$$x \succsim y \Leftrightarrow u(x) \geq u(y). \tag{1}$$

However, completeness is not a precondition for utility maximization. Maximizing a utility function u is consistent with an underlying preference relation

[1] Some authors use strict preferences $\succ$ as the primitive and define weak preference $\succsim$ and indifference $\sim$ from $\succ$ (see e.g. Fishburn [9] for details). In a revealed preference context, weak preference is often the natural starting point, since if x were chosen when y was also available, x and y may or may not be indifferent.

$\succsim$ if dominated alternatives are not selected, i.e. if

$$x \succ y \Rightarrow u(x) > u(y). \tag{2}$$

Given X and $\succsim$, there exists a utility function u satisfying (2) if and only if strict preference $\succ$ is acyclic (i.e. there is not a finite t and a sequence $x_1, x_2, \ldots, x_t$ such that $x_1 \succ x_2 \succ \cdots \succ x_t \succ x_1$), cf., e.g., Fishburn [11].

We have therefore two interpretations of a utility representation, (1) and (2), depending on whether completeness is assumed or not. In order to obtain a representation (1), transitivity must hold, while in (2), acyclicity must hold. Transitivity implies acyclicity but the converse is not true. Intransitivity and cycles are however closely related preference phenomena and in practice one often seeks to find violations of transitivity by means of cycles (see, e.g., [25]). Thus cycles (or intransitivity for that matter), not incompleteness, is the phenomenon of interest here.

The axiom of completeness only plays a role if, when imposing it, we can point to intransitivities but not to cycles. To illustrate, suppose that there are three alternatives $X = \{x, y, z\}$ and $x \succ y$, $y \succ z$ and $x \not\succ z$. Then if completeness is assumed we must necessarily have $z \succsim x$ and there is no utility representation in the sense of (1) due to the implied intransitivity of $\succsim$. On the other hand, without completeness $x \not\succ z$ may indicate that x and z cannot be compared with $\succsim$ and in this case a utility representation in the sense of (2) exists.

In relation to tests of the utility maximizing hypothesis in discrete choice experiments there is consequently no particular reason to assume that choice observations are drawn from an underlying complete ordering rather than from a partial ordering on the alternatives actually compared. For the purpose of testing for the existence of a utility representation it suffices to test for cycles or intransitivity within the observed choices. Whether or not the axiom of completeness is assumed is more or less a matter of taste.

2.1 What is tested in recent studies?

In light of the very limited role of the axiom of completeness it may seem surprising that recent papers are preoccupied with testing whether completeness is satisfied or not in choice experiments, see, e.g., [41] and [35] (see also [33] and [42]). Apparently, it is because experimenters find that there is a risk that respondents when confronted with various alternative options (that are difficult to grasp as for example in case of health care interventions or environmental evaluation) have no well-formed preferences, but still try to deliver an answer in order not to disappoint the experimenter. Such behavior is then assumed to be revealed by conflicting rankings in case of repeated choice — taken as a sign of incompleteness.

But what is in fact tested in these studies? Ryan and San Miguel [35], for example, develop a test for completeness interpreted as the assumption that individuals have what they call "well-defined preferences" for any choice they are presented to. In the experiment, two specific choice situations, choice A and choice B, were both repeated during the experiments, with choice A repeated before choice B was introduced (and then later repeated).[2] In each choice situation, two alternatives a and b were presented and the respondent was asked to select one of the following options: 1) "Strongly prefer a" 2) "Prefer a" 3) "Indifferent" 4) "Prefer b" 5) "Strongly prefer b". If no reversals in stated preferences neither in the second round of choice A nor in the second round of B was observed, Ryan and San Miguel interpret this as (an indication of) "complete preferences".[3] If preference reversals occurred both in A and B this was interpreted as "incomplete preferences". Preference reversal in A but not in B was interpreted as a "learning effect", and, finally, if there was a preference reversal in B but not in A then the interpretation was "random error" (or "tiredness").

[2] The whole procedure was then again repeated in three waves.

[3] Certain changes in stated preferences, such as a change from "strongly prefer a" to "prefer a", was not counted for a preference reversal.

We may try to get an intuitive grasp of the preference phenomena discussed by Ryan and San Miguel and others using Figure 1 (see below), where we have also added a fourth possibility, "unstable preferences". Figure 1 illustrates for simplicity a *binary* choice situation with $X = \{x, y\}$ where the choice situation is repeated a number of times. In Figure 1A illustrates a situation where there is a preferred alternative x but random shocks occasionally change observed choice (can be called "random error"). Figure 1B illustrates a situation, where x and y cannot be meaningfully compared and the choices - forced through by the analyst - are arbitrary ("incomplete preferences"). In Figure 1C there is a learning effect in the sense that preferences seemingly converge after initial randomness ("learning").[4] Finally, in Figure 1D we have illustrated another possibility, preferences are "complete" but change over time ("unstable preferences").

As Figures 1A and 1B illustrates, there is no point in distinguishing between incompleteness and random error since indecisiveness and noise cannot be disentangled based on such choice observations.[5] Hence, if the underlying model is assumed to be a random preference/utility model (which seems sensible provided that "mistakes" are to be expected in all choice experiments) incompleteness cannot be separated from noise — this will be further discussed in Section 4. Learning effects, on the other hand, are quite different due to the fact that choices become more stable over time, i.e. noise is reduced over time. By repeating a choice once we cannot distinguish between learning and measurement error. By repeating more than once we can observe if stated preference seems to converge, see Section 4. Unstability, as in Figure 1D, could be identified with positive autocorrelation between

[4]The word "learning" may be imprecise since learning (in the sense of becoming more well-informed or wiser) is not necessarily the same as convergence of choice. Choices may, for example, initially be stable due to ignorance and gradual learning about the true complexity of the matter (or preference for diversity) may introduce doubt — and thereby unstability.

[5]Note that by chosing from the same choice sets twice, preference reversals cannot be explained by "menu-dependent" choice rules, see, e.g., Sen [40].

successive choices, but such an effect takes more than a single repetition to disentangle from other types of effects. In particular, testing the difference between learning and unstability is impossible using tests as in [35].[6]

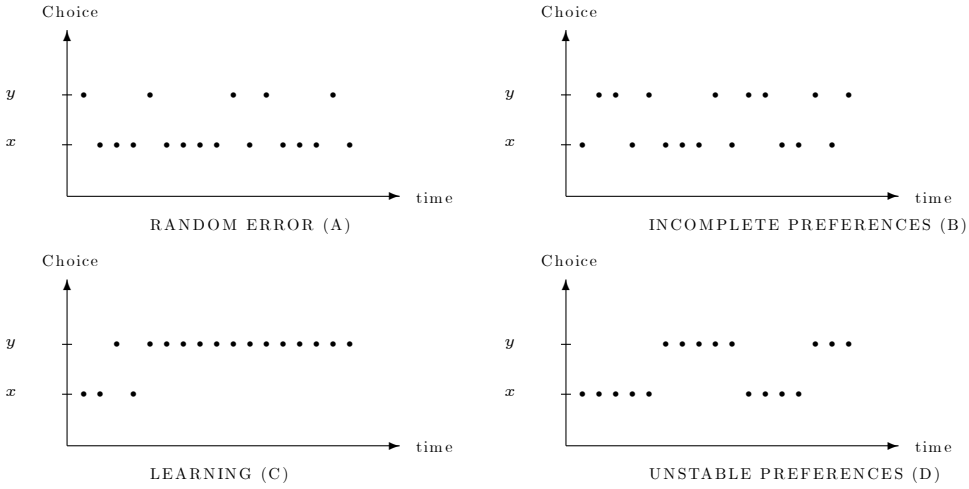

Figure 1

Completeness interpreted as having well-formed preferences cannot be accepted or rejected on the basis of a standard questionnaire study, because the results of such a study will always be reported in terms of relative frequencies. If preference for x over y is defined as "in a majority of cases x is preferred to y", any two alternatives can be compared. The only exception is the case where respondents can refuse to answer a question. Therefore, it might be an idea as, e.g., suggested in Oliver [33], to add to each question

[6]The impossibility of making a clear distinction between incompleteness and noise is, in fact, very well illustrated by the empirical examples in [35]. One of the examples involves a set of questions concerning supermarket attributes, where the alternatives are only vaguely specified. For instance, prices can be "high, medium or low", without any clear quantitative specification. Obviously, many respondents will react to this by simply refusing to answer, or — as a more polite alternative — to give only vague answers. It is really a matter of taste whether this should be taken as an indication of incompleteness, an indication of noise, or an indication of alternatives that are difficult to distinguish from each other. Not surprisingly the number of imprecise preferences in the supermarket study is, in most cases, higher than in the two other studies presented, where the description of the alternatives is more precise.

a response category labelled "comparison meaningless". If all respondents put their votes in that category we can, with some weight, conclude that either the ordering is incomplete, or the alternatives are so vaguely defined that the respondents are unable to answer. However, this category must certainly not be confused with the mid–category labelled "indifferent", which may very well be selected as the result of a careful comparison of well–defined alternatives.

3 Testing transitivity

Since inconsistency in binary choices essentially is related to intransitivities (or cycles) we shall now discuss whether tests for transitivity can be performed. Assume, for simplicity, that we have a data set of pairwise comparisons for a given respondent facing repeated choices between alternatives in a given set X, where each question has the form "which of the following two alternatives x and y do you prefer?".[7] Consequently, we make the simplifying assumption of unambiguous answers, i.e. answers like "I don't know" or "I am indifferent" are not accepted.[8]

Now, let $p(x|x,y)$ be the probability that x is chosen among alternatives x and y. Since indifference is not possible, we have

$$p(x|x,y) = 1 - p(y|x,y).$$

Under our simplifying assumptions, the probabilities $p(x|x,y)$ can be estimated by the corresponding relative frequencies $\hat{p}(x|x,y)$. The preference

[7] It is easy to generalize the method presented here to the case where three or more alternatives are presented in each comparison, see, e.g., Block and Marschak [3].

[8] As we shall discuss in Section 4, questions allowing for indifference can, in a more structured statistical model where transitivity is an inherent property, be handled simply by exclusion of the indifference–answers from the data set.

relation induced by the choice probabilities is given by

$$x \succsim y \Leftrightarrow p(x|x,y) \geq \frac{1}{2},$$

and the estimate of this relation becomes, accordingly,

$$x \hat{\succsim} y \Leftrightarrow \hat{p}(x|x,y) \geq \frac{1}{2}\,.$$

Since any pair (x,y) of alternatives will satisfy either $x \succsim y$ or $y \succsim x$, a test for completeness is meaningless. Of course, it can be argued that if $\hat{p}(x|x,y)$ is close to 1/2 it is an indication of "coin–flip" answers, which could be explained by the respondents' lack of ability to perform a relevant comparison. But it can also be taken, simply, as an indication of x and y being very similar alternatives, and there is no way of distinguishing between these two explanations.

However, a test for transitivity is possible. Transitivity of the induced relation $\succsim$ means (ignoring ties $p(x|x,y) = \frac{1}{2}$ which are not likely to occur) that for any triple (x,y,z) we have[9]

$$[p(x|x,y) < \frac{1}{2} \;\text{ and }\; p(y|y,z) < \frac{1}{2}] \;\Rightarrow\; p(x|x,z) < \frac{1}{2}.$$

Thus, in order to test that a *given* triple (x,y,z) does not give rise to any violation of transitivity, the following procedure will suffice:

First, check whether the three comparisons (x,y), (y,z) and (x,z) all result in *significantly decisive* conclusions. Or, equivalently, check by standard binomial tests that all three estimates $\hat{p}(x|x,y)$, $\hat{p}(y|y,z)$ and $\hat{p}(x|x,z)$ are significantly different from $\frac{1}{2}$. If this is not the case, transitivity must necessarily be accepted as far as this triple is concerned. If the three comparisons are decisive, check whether their ordering is in accordance with transitivity

[9] Or, equivalently, *lack of transitivity* means that there exists a triple (x,y,z) (think of it as a 3–cycle $x \to y \to z \to x$) for which the three probabilities $p(x|x,y)$, $p(y|y,z)$ and $p(z|z,x)$ are all less than $\frac{1}{2}$. Fishburn [10] surveys theories of stochastic transitivity.

or not. If not (i.e. if the ordering is cyclic, which happens in 2 of the 8 possible cases), transitivity is rejected, otherwise it must be accepted.

An overall test for transitivity is, in principle, just a matter of doing this for all possible triples. But here we must (in particular if many alternatives are involved) take "mass significance" into account, i.e. the phenomenon that when many tests on level (say) 95% are performed, some of them will usually be significant just by accident. A procedure that takes this into account goes as follows:

First, isolate all pairs (x, y) for which $\hat{p}(x|x, y)$ is significantly different from $\frac{1}{2}$ on a suitable level. This level should be determined in such a way that we are almost certain that *all* these "decisive" comparisons are actually correct. Since there are $\binom{n}{2}$ pairs of alternatives (where n is the number of alternatives in X), the only way of ensuring this is to perform the tests on level $1 - \alpha/\binom{n}{2} = 1 - 2\alpha/(n(n-1))$, where α is chosen (as usual) to be 0.05 or 0.01 or whatever is preferred. In this way we can be sure that a "false decisive comparison" occurs with probability at most α. When these pairs have been isolated, check that the corresponding graph has no cycles. If there are cycles, transitivity is rejected, otherwise it cannot be rejected.

It is possible to invent more refined versions of this test. Iverson and Falmagne [16] have a quite complicated study of the distributions of the statistics related to this kind of hypothesis testing. But their basic idea is the same, and we believe that the simple procedure proposed here will work satisfactory in practice, at least when the number of alternatives is moderate. The test suggested here is also related to a test for transitivity applied by Tversky [46] in an experimental study, where the alternative to transitivity was more narrowly specified by the nature of the experiment. Apart from this, the procedure suggested here seems not to have been mentioned in the literature before (see, however [3] and [17] for studies of related problems). The reason for this is probably that the more structured statistical models (like the random utility model) that are usually applied in this context have

transitivity as an intrinsic property. Thus, the acceptance of one of the models discussed in Section 4 (e.g. by an ordinary goodness–of–fit test) is an implicit acceptance of transitivity.

A data set of pairwise comparisons could also be collected from a *group* of respondents and we can imagine that the experimenter wants to test if it is possible to interpret the choices as if they were all generated by a single "representative individual" with stochastic transitive preferences. In this case it is more likely that a sufficiently large data set is available for statistical testing. Problems of this kind might have interest from, e.g., a social choice perspective. It is well-known, however, that transitivity of individual preferences do not necessarily imply a transitive preference relation for the representative individual and vice versa (e.g. May [25]).

3.1 Recent experimental studies

In recent literature on choice experiments, some studies, e.g., Sælensminde [44][45], state that the transitivity axiom is tested using so-called Ray diagrams but it is not explicitly defined what is meant by transitivity, and it is questionable whether the consistency tests performed relate to a test of this axiom or other preference phenomena (such as linearity). Other studies, e.g., Carlsson and Martinsson [5] and McIntosh and Ryan [32] use a random utility model but seem to disregard the stochastics when they test for transitivity with the purpose of providing internal validation of their framework.

These studies are concerned with the absolute or relative number of respondents that show some sort of intransitive behavior in at least one occasion. If the number of comparisons performed by each respondent is large, and if the alternatives are difficult to distinguish, many of the respondents are likely to get into some sort of self-contradictory behavior. But this does not necessarily imply that there is an underlying preference relation which is intransitive.

In relation to these tests of preference axioms, recent experimenters have

also been concerned with the idea of exclusion of data from respondents (for further analysis data analysis) if the entire pattern of those respondents choices do not pass all "consistency checks". The choice models suggest that there is no reason whatsoever to eliminate data, as long as the observed violations are within the range of what could be expected in the relevant statistical model: Whereas experimenters should be free to interpret single deviations from "consistency checks" as they want in its specific context, it would be wrong to associate such consistency checks with validations of the relevant discrete choice models or the use of such models for utility assessments in applied welfare studies.

4 On testing preference axioms (and detecting other preference phenomena) in the random utility framework

In this section we discuss how concepts like incompleteness, learning, tiredness, and related issues can be analyzed in the framework of more structured statistical models. For a general treatment of such models, see, e.g., McFadden [29].

Discrete comparisons, in this context, refers to a situation where a number of respondents are confronted with a number of questions of the form "which of the following k alternatives do you prefer".[10]

The classical model for this kind of situations is the so-called Bradley–Terry model (see [4]), which can be stated as follows: Let π_x denote the (more or less fictive) probability that a (random) respondent, when presented to the entire set $X = \{1, \ldots, n\}$ of alternatives, answers "x". Thus, $\pi_1 + \cdots + \pi_n = 1$, provided that an answer must be given, which is assumed for the moment.

A crucial (and in some contexts questionable) assumption, called the

[10] The case $k = 2$ corresponds to the pairwise comparison setup of the previous section.

"axiom of independence of irrelevant alternatives"[11], is that if only a subset of the set X of alternatives is presented to the respondent, then the probabilities can be derived from the situation involving the full set of alternatives as the conditional probabilities, given that the choice happens to fall in the subset.[12] For example, if three alternatives x, y and z are presented, we have (with an obvious extension of the notation used in Section 4)

$$p(x|x,y,z) = \frac{\pi_x}{\pi_x + \pi_y + \pi_z}.$$

A nice property of this model, which relates to our previous discussion of completeness, is that it is consistent with the simplest possible handling of "don't know" answers, in the following sense. If an indifference category – which can suitably be named 0 – is added, and if we can rely on the assumption that this alternative plays a role which is similar to any other alternative, then the "don't know" answers can be handled simply by removing them from the data set. For example, if two alternatives x and y are presented, the probability of choosing x when indifference is allowed becomes

$$p(x|x,y,0) = \frac{\pi_x}{\pi_x + \pi_y + \pi_0}\,.$$

But the conditional probability of selecting x, given that either x or y is selected becomes

$$p(x|x,y) = \frac{\pi_x}{\pi_x + \pi_y}$$

which according to the assumption coincides with the probability of selecting x when 0 is not among the alternatives presented.

Another nice property is automatic transitivity of the induced prefer-

[11] See, e.g., Luce [22], Marschak [23] and McFadden [28].

[12] The drawback of this assumption is that one can easily invent examples where it is unrealistic. If a pair $\{x, y\}$ of clearly distinct alternatives is extended by an alternative z which appears very similar to x, then it is not likely that the probability of selecting y will become much smaller — though this is actually what the formula suggests. Nevertheless, the model has been widely applied

ence relation. Indeed, since $x \prec y$ is obviously equivalent to $\pi_x < \pi_y$, the transitivity condition reduces to the trivial statement

$$\pi_x < \pi_y \text{ and } \pi_y < \pi_z \Rightarrow \pi_x < \pi_z \,.$$

This result is further supported by the result noticed by McFadden [28] that the Bradley–Terry model can be derived as a random utility model, i.e. a model that explains the choice made by a respondent as the one that maximizes the utility over the alternatives presented. Since choices vary from occasion to occasion and between respondents, this utility function has to be random.[13] More specifically, let v be a function which to each alternative $x \in X$ assigns a real number $v(x)$, which can be interpreted as a sort of "average utility" in the population. The random utilities determining the choices are assumed to take the form

$$U_{ri}(x) = v(x) + \varepsilon_{xri},$$

where ε_{xri} is a random variable associated with alternative x in the i'th choice performed by respondent r. These "error terms" are assumed to be independent and identically distributed, and the choice made by a respondent in any choice situation is assumed to be the choice that maximizes the value of the random utility function U_{ri}.[14]

Falmagne [8], McFadden and Richter [31] and others have established necessary and sufficient conditions for theoretical choice probabilities to be

[13] The *random preference model* offers a quite distinct stochastic specification, see, e.g., Loomes and Sugden [20].

[14] What McFadden [28] showed was that if the common distribution of the ε_{xri} is assumed to be the normalized extreme value distribution (c.d.f. $P(\varepsilon_{xri} \leq z) = \exp(-\exp(-z))$), then this model coincides with the Bradley-Terry model with parameters

$$\pi_{x_i} = \frac{\exp(v(x_i))}{\exp(v(x_1)) + \cdots + \exp(v(x_k))},$$

for alternatives $X = \{x_1, \ldots, x_k\}$.

consistent with random utility maximization. McFadden [30] contains many references. Stochastic tests, however, have been much less developed; albeit, see Koning and Ridder [18] for a related study.

4.1 Tests of preference phenomena

There is an extensive literature on tests of various types of respondent behavior in a random utility framework, and we shall refrain from an attempt to survey this broad field (see, however, Hensher et al. [13], Louviere [21], and Swait and Adamowicz [43] for other recent discussions). Rather, we shall try to synthesize how some of the previously mentioned, and somewhat vaguely defined, preference phenomena related to discrete comparisons can be formalized and tested within the framework of the random utility model.

The fact that *incompleteness* is indistinguishable from close similarity of alternatives is clearly demonstrated by the model when a scale parameter is introduced for the error term of the random utility function. If we write the random utility as

$$U_{ri}(x) = v(x) + \sigma\varepsilon_{xri}$$

where σ is a scale parameter, similar to the standard deviation in a regression model (ε_{xri} is still assumed to be normalized extreme value distributed), it becomes clear that a large degree of incompleteness (meaning that respondents seem to give their answers more or less at random) is equivalent to a large value of σ, whereas close similarity of alternatives means that the values $v(x)$ all lie in some narrow interval. But it is well-known that an upscaling of the function v is equivalent to a downscaling of the error term ε_{xri}, and vice versa, and it is an intrinsic property of this model that it cannot distinguish between these two phenomena. To avoid this overparametrization we may as well take $\sigma = 1$ in the model where σ is constant (see, e.g., Ben-Akiva and Lerman [2]).

In addition, the idea of a scale parameter on the error term allows us to

build a *learning effect* into the model in the following way. If respondents are exposed to the same set of alternatives several times, or to different combinations involving the same alternatives, a learning effect may be interpreted as respondents becoming more and more stable and consistent in their selections. This phenomenon becomes possible in the model if we allow for a scale parameter σ_i that varies from occasion to occasion (i). If σ_i decreases, a learning effect is present. The phenomenon that σ_i increases at some point seems to be appropriately described by the word *tiredness*.

Heterogeneity between respondents can also be modelled. In practice, a realistic expectation is that the random variation from respondent to respondent is more pronounced than the variation from occasion to occasion for the same respondent. Moreover, a specific respondent may very well show a stable behavior which is different from that of another respondent. The deterministic utility function $v(x)$ represents a kind of population average, but respondents may have individual preferences that are different from this average. A model that takes this into account could be a variance–component–type model based on a random utility function of the form

$$U_{ri}(x) = v(x) + \omega\delta_{xr} + \sigma\varepsilon_{xri}$$

where $\omega\delta_{xr}$ is an error term of the same kind as $\sigma\varepsilon_{xri}$, except that it is specific to the alternative x and the respondent r, but independent of the occasion i. Computationally, this model is difficult to handle, but conceptually this is exactly what is needed to describe heterogeneity between respondents. This model is not equivalent to a Bradley–Terry model or any other simple model. Models of this kind are usually specified with normal rather than extreme value distributed error terms.

An *indifference category* can be incorporated in the model in a simple way, which in most cases is likely to be more realistic than the ignorance–of–indifference–cases method proposed earlier. Consider for simplicity the case

of pairwise comparisons. Instead of assuming

$$\text{Choice} = \begin{cases} x & \text{if } U_{ri}(x) > U_{ri}(y) \\ y & \text{if } U_{ri}(x) < U_{ri}(y) \end{cases}$$

we could assume, for some parameter $\beta_0 > 0$ which can be interpreted as the "least noticeable utility difference", that

$$\text{Choice} = \begin{cases} x & \text{if } U_{ri}(x) > U_{ri}(y) + \beta_0 \\ 0 & \text{if } |U_{ri}(x) - U_{ri}(y)| \leq \beta_0 \\ y & \text{if } U_{ri}(x) < U_{ri}(y) - \beta_0. \end{cases}$$

One might even consider models where the parameter β_0 varies from respondent to respondent, in accordance with the fact that some people are more hesitant with decisive conclusions than others. A similar model — with an additional parameter $\beta_1 > \beta_0$ to determine the threshold between "preference" and "strong preference" — can be used in situations where the responses are given on (say) a five–point scale (as, e.g., in [35]). These models are closely related to the models for discrete ordinal data described in McCullagh [27].

As a final remark we mention the possibility of taking a *preference–for–first–met–alternative* parameter into the model. The order in which alternatives are presented may influence the decision taken, typically by giving a higher probability to the alternatives presented first. For this reason, it is important to balance the questionnaires in such a way that the alternatives presented are not always given in the same order. Provided that this has been done, there is a rather straightforward way of building this into the model. In the case of pairwise comparisons, it can be done by the introduction of an extra preference–for–first–met–parameter, which is simply added to the deterministic utility function's value for the first alternative before the maximization. If the utility function is written as a linear combination of covariate values, this has the simple interpretation that the property of

being presented first is an extra measure of quality (represented by a dummy covariate) with potential (positive or negative) influence on the choice. For triplewise comparisons and higher, it becomes a bit more complicated.

5 Conclusion

In this paper we have examined the possibilities for embedding tests of preference axioms within probabilistic choice models. We have in particular discussed the role of completeness and transitivity, and provided some suggestions for dealing with notions like learning or tiredness, heterogeneity, indifference categories and ordering effects within the random utility model. It has been argued that both completeness and transitivity in theory play a role but the empirical relevance of particularly the completeness axiom is strongly limited. Transitivity can be tested within a frequency of choice model, although for most realistic data sets it seems unlikely that transitivity can be rejected at the individual level. In this respect it seems reasonable to work with statistical models that treat these properties as inherent.

Having said that, it should be emphasized that we do not in any respect want to downplay the importance of internal and external validations of discrete choice models. Although statistical test are sometimes meaningless in relation to specific models, critical investigations of respondents' ability to deliver meaningful, reliable, and useful answers in stated preference experiments are no less needed.

References

[1] Afriat, S.N., 1967, The construction of utility functions from expenditure data. *International Economic Review* **8**, 57-77.

[2] Ben-Akiva, M.E., Lerman S.R., 1985, *Discrete choice analysis: theory and application to travel demand*. MIT Press, Cambridge, Mass.

[3] Block, H.D., Marschak, J., 1960, Random orderings and stochastic theories of responses. In *Contributions to probability and statistics*, Olkin et al. (Eds.), Stanford University Press.

[4] Bradley, R.A., Terry, M.E., 1952, Rank analysis of incomplete block designs. *Biometrika* **39**, 324-345

[5] Carlsson, F., Martinsson, P., 2001, Do hypothetical and actual marginal willingness to pay differ in choice experiments? *Journal of Environmental Economics and Management* **41**, 179-192.

[6] Cox, J.C., 1997, On Testing the Utility Hypothesis. *Economic Journal* **107**, 1054-78.

[7] Diewert, W.E., 1973, Afriat and revealed preference theory. *Review of Economic Studies* **40**, 419-425.

[8] Falmagne, J., 1978, A representation theorem for finite random scale systems. *Journal of Mathematical Psychology* **18**. 52-72.

[9] Fishburn, P., 1970, *Utility theory for decision making.* John Wiley & Sons, Inc.

[10] Fishburn, P., 1999, Stochastic utility theory, Ch. 7 in Barberà, Salvador; Hammond, Peter; Seidl, Christian (Eds.). *Handbook of Utility Theory.* Kluwer.

[11] Fishburn, P., 1999, Preference structures and their numerical representations. *Theoretical Computer Science* **217**, 359-383.

[12] Fleissig A.R., Whitney, G.A., 2005, Testing for the significance of violations of Afriat's inequalities. *Journal of Business and Economic Statistics* **23**, 355-362.

[13] Hensher , D., Louviere, J., Swait, J., 1999, Combining source of preference data. *Journal of Econometrics* **89**, 197-221.

[14] Hougaard, J.L., Tjur, T., Østerdal, L.P., 2004, The role of preference axioms and respondent behaviour in statistical models for discrete choice. *Preprint*, Center for Statistics, Copenhagen Business School.

[15] Houthakker, H.S., 1950, Revealed preference and the utility function. *Economica* **17**, 159-174.

[16] Iverson, G., Falmagne, J.-C., 1985, Statistical issues in measurement. *Mathematical Social Sciences* **10**, 131-153.

[17] Kendall, M.G., Babington Smith, B., 1940, On the method of paired comparisons. *Biometrika* **31**, 324-345

[18] Koning, RH., Ridder, G., 2003, Discrete choice and stochastic utility maximization. *Econometrics Journal* **6**, 1-27

[19] Johnson. F.R., Mathews, K.E., 2001, Sources and effect of utility-theoretic inconsistency in stated-preference surveys. *American Journal of Agricultural Economics* **84**, 1328-1333.

[20] Loomes, G., Sugden, G., 1995, Incorporating a stochastic element in decision theories. *European Economic Review* **39**, 641-648.

[21] Louiviere, J.J., 2001, What if consumer experiments impact variances as well as means? Response variability as a behavioral phenomenon. *Journal of Consumer Research* **28**, 506-511.

[22] Luce, R.D., 1959, *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons.

[23] Marschak, J., 1960, Binary-choice constraints on random utility indicators. In Arrow, K., Karlin, S., Suppes, P. (Eds.), *Mathematical Methods in the Social Sciences*. Stanford University Press.

[24] Matzkin, R.L., Richter M.K., 1991, Testing strictly concave rationality. *Journal of Economic Theory* **53**, 287-303.

[25] May, K., 1954, Intransitivity, utility, and the aggregation of preference patterns. *Econometrica* **22**, 1-13.

[26] Matzkin, R., 1991, Axioms of revealed preference for non-linear choice sets. *Econometrica* **59**, 1779-1786.

[27] McCullagh, P., 1980, Regression models for ordinal data. *Journal of the Royal Statistical Society B* **42**, 109–142

[28] McFadden, D., 1973, Conditional logit analysis of qualitative choice behavoir. In Zarembka, P. (Ed.) *Frontiers in Econometrics*, Academic Press.

[29] McFadden, D., 1986, The choice theory approach to market research. *Marketing Science* **5**, 275-297.

[30] McFadden, D., 2005, Revealed stochastic preference: a synthesis. *Economic Theory* **26**, 245-264.

[31] McFadden, D., Richter, M.K., 1991, Stochastic Rationality and Revealed Stochastic Preference, in J. Chipman, D. McFadden, and M.K. Richter (eds) *Preferences, Uncertainty and Rationality*. Westview Press.

[32] McIntosh, E., Ryan, M., 2002, Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: Implications of discontinuous preferences. *Journal of Economic Psychology* **23**, 367-382.

[33] Oliver, A., 2000, Complete preferences over health states: A reply to the paper by Shiell *et al.*. *Health Economics* **9**, 727-728.

[34] Richter, M. K., 1966, Revealed preference theory. *Econometrica* **34**, 635-645.

[35] Ryan, M., San Miguel, F., 2003, Revisiting the axiom of completeness. *Health Economics* **12**, 295-307.

[36] Ryan, M., Bate, A., 2001, Testing the assumptions of rationality, continuity and symmetry when applying discrete choice experiments in health care. *Applied Economics Letters* **8**, 59-63.

[37] Ryan, M., McIntosh, E., Shackley, P., 1998, Methodological issues in the application of conjoint analysis in health care. *Health Economics* **7**, 373-378.

[38] San Miguel, F., Ryan, M., Scott, A., 2002, Are preferences stable? The case of health care. *Journal of Economic Behavior and Organization* **48**, 1-14.

[39] Scott, A., 2002, Identifying and analysing dominant preferences in discrete choice experiments: An application in health care. *Journal of Economic Psychology* **23**, 383-398.

[40] Sen, A., 1993, Internal consistency of choice. *Econometrica* **61**, 495-521.

[41] Shiell, A., Seymour, J., Hawe, P., Cameron, S., 2000, Are preferences over health states complete? *Health Economics* **9**, 47-55.

[42] Shiell, A., Seymour, J., Hawe, P., Cameron, S., 2000, Will our understanding of completeness ever be complete? *Health Economics* **9**, 729-731.

[43] Swait, J., Adamowicz, W., 2001, Choice environment, market complexity, and consumer behavior: A theoretical and emperical approach for incorporating decision complexity into models of consumer choice. *Organizational Behavior and Human Decision Processsses*, **86**, 141-167.

[44] Sælensminde, K., 2001, Inconsistent choices in stated choice data. *Transportation* **28**, 269-296.

[45] Sælensminde, K., 2002, The impact of choice inconsistencies in stated choice studies. *Environmental and Resource Economics* **23**, 403-420.

[46] Tversky, A., 1969, Intransitivity of preferences. *Psychological Review* **76**, 31-48.

[47] Varian, H.R., 1982, The nonparametric approach to demand analysis. *Econometrica* **50**, 945-973.